
Supplementary Material: Whole-Body-Conditioned Ego-Centric Video Prediction

Anonymous Author(s)

Affiliation

Address

email

1 The structure of the Appendix is as follows: we start by describing how to apply PEVA for more
2 planning ability with in Section 1, and then include more qualitative results in Section 2, including
3 atomic actions, simulate counterfactuals, and long video generation.

4 1 Planning with PEVA

5 Here we describe how to use a trained PEVA to plan action sequences to achieve a visual target. We
6 formulate planning as an energy minimization problem and perform standalone planning in the same
7 way as NWM (Bar et al., 2024) using the Cross-Entropy Method (CEM) (Rubinstein, 1997) besides
8 minor modifications in the representation and initialization of the action.

9 For simplicity, we conduct two experiments where we only predict moving either the left or right
10 arm controlled by predicting the relative joint rotations represented as euler angles. For each re-
11 spective arm we control only the shoulder, upper arm, forearm, and hand leaving our actions space
12 as $4 \times 3 = 12$ where we have $(\Delta\phi_{\text{shoulder}}, \Delta\theta_{\text{shoulder}}, \Delta\psi_{\text{shoulder}}, \dots, \Delta\phi_{\text{forearm}}, \Delta\theta_{\text{forearm}}, \Delta\psi_{\text{forearm}})$.
13 We initialize mean $(\mu_{\Delta\phi_{\text{shoulder}}}, \mu_{\Delta\theta_{\text{shoulder}}}, \mu_{\Delta\psi_{\text{shoulder}}}, \dots, \mu_{\Delta\phi_{\text{forearm}}}, \mu_{\Delta\theta_{\text{forearm}}}, \mu_{\Delta\psi_{\text{forearm}}})$ and variance
14 $(\sigma_{\Delta\phi_{\text{shoulder}}}^2, \sigma_{\Delta\theta_{\text{shoulder}}}^2, \sigma_{\Delta\psi_{\text{shoulder}}}^2, \dots, \sigma_{\Delta\phi_{\text{forearm}}}^2, \sigma_{\Delta\theta_{\text{forearm}}}^2, \sigma_{\Delta\psi_{\text{forearm}}}^2)$ as the mean and variance of the next
15 action across the training dataset for these segments.

Table 1: Mean and Variance of relative rotation as euler angles (ϕ, θ, ψ) for arm segments computed across the training dataset.

Segment	Statistic	Right Arm	Left Arm
Shoulder	Mean	(0.0027, -0.0012, -0.0015)	(0.0624, 0.0687, 0.1494)
	Variance	(0.0010, -0.0006, 0.0003)	(0.0625, 0.0697, 0.1496)
Upper Arm	Mean	(0.0107, -0.0011, -0.0020)	(0.1119, 0.1647, 0.1791)
	Variance	(-0.0062, -0.0004, -0.0013)	(0.0991, 0.1593, 0.1611)
Forearm	Mean	(0.0068, -0.0035, 0.0077)	(0.1937, 0.2107, 0.2261)
	Variance	(-0.0036, -0.0063, 0.0002)	(0.1791, 0.2012, 0.2186)
Hand	Mean	(0.0065, 0.0001, 0.004,)	(0.2417, 0.229, 0.2631)
	Variance	(-0.0024, -0.0032, -0.0001)	(0.2126, 0.2237, 0.2475)

16 We assume the action is a straight continuous motion. Thus we repeat this action for our sequence
17 length, in our case $T = 8$ and optimize the delta actions. The time interval between steps is fixed at
18 $k = 0.25$ seconds. All other hyperparameters remain the same as in NWM (Bar et al., 2024).

19 1.1 Qualitative Planning Results

20 Due to time constraints, we focus our investigation on arm movements—arguably the most complex
 21 among body actions. While this remains an open problem, we present preliminary results using
 22 PEVA with CEM for standalone planning. This setting simplifies the high-dimensional control space
 23 while still capturing key challenges of full-body coordination.

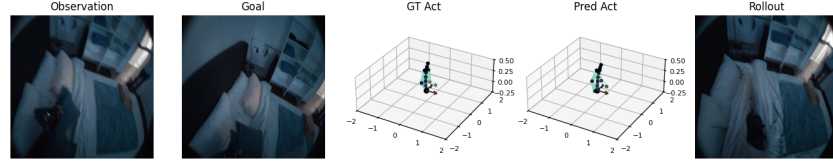


Figure 1: In this case, we are able to predict a sequence of actions that pulls our left arm in, similar to the goal.

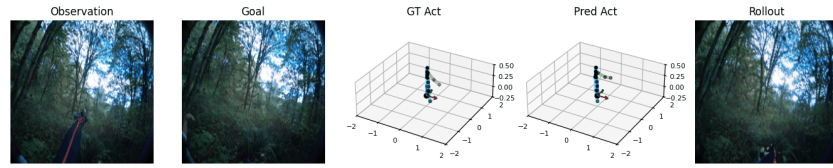


Figure 2: In this case, we are able to predict a sequence of actions that lowers our left arm, but not the same amount as the groundtruth sequence as we can see in the pose and hand at the bottom of our rollout.

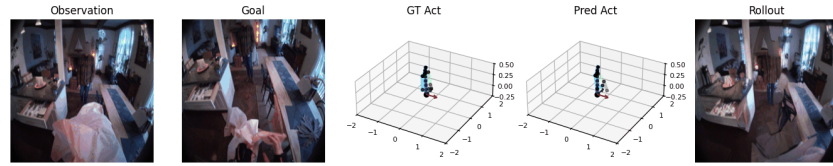


Figure 3: In this case, we are able to predict a sequence of actions that lowers our left arm that lowers the tissue. However, the goal image still has the tissue visible.

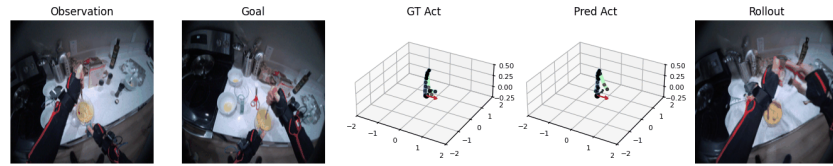


Figure 4: In this case, we are able to predict a sequence of actions that raises our right arm to the mixing stick. We see a limitation with our method as we only predict the right arm so we do not predict to move the left arm down accordingly.

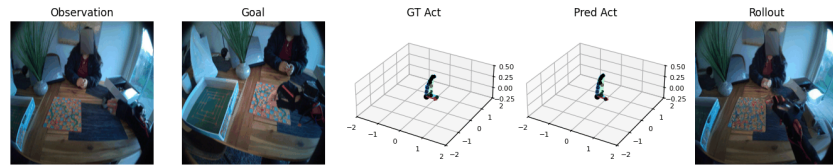


Figure 5: In this case, we are able to predict a sequence of actions that moves our right arm toward the left but not quite enough. We see a limitation with our method as we only predict the right arm so we do not predict any necessary additional body rotations.

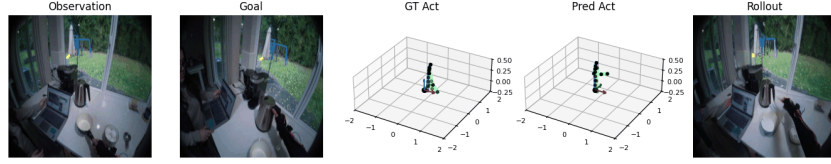


Figure 6: In this case, we are able to predict a sequence of actions that reaches toward the kettle but does not quite grab it as in the goal.

2 Qualitative Results

In the main paper, we provide three types of visualization: PEVA can simulate counterfactuals, generate videos of atomic actions, and long video generation.

Here, we show more qualitative results following the settings in main paper:

Counterfactuals: We start by sampling multiple action candidates and simulate each action candidate using PEVA via autoregressive rollout. Finally, we rank each action candidate’s final prediction by measuring LPIPS similarity with the goal image.

Atomic Actions: We decompose complex motions into atomic actions. By analyzing joint trajectories over short windows, we extract video segments exhibiting fundamental movements.

Long Video Generation: We evaluate the model’s ability to maintain visual and semantic consistency over extended prediction horizons. PEVA generates coherent 16-second rollouts conditioned on full-body motion.

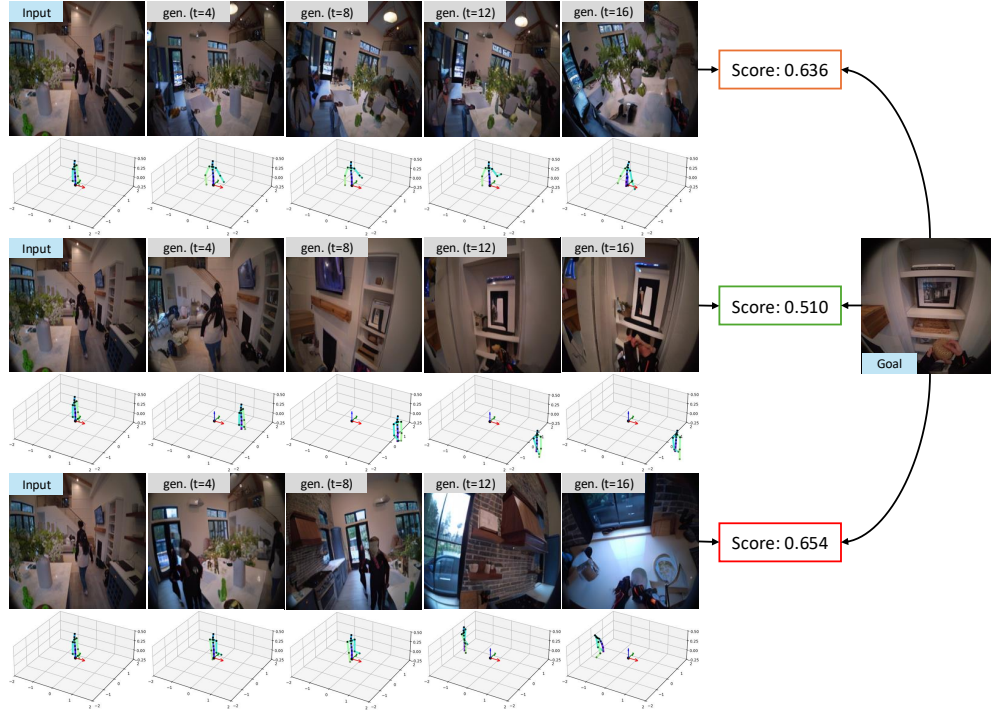


Figure 7: **Planning with Counterfactuals.** We demonstrate a planning example by simulating multiple action candidates using PEVA and scoring them based on their perceptual similarity to the goal, as measured by LPIPS (Zhang et al., 2018). PEVA enables us to rule out action sequences that grab the nearby plants and go to the kitchen and mix ingredients. PEVA allows us to choose the most correct action sequences that grabs the box from the shelf.

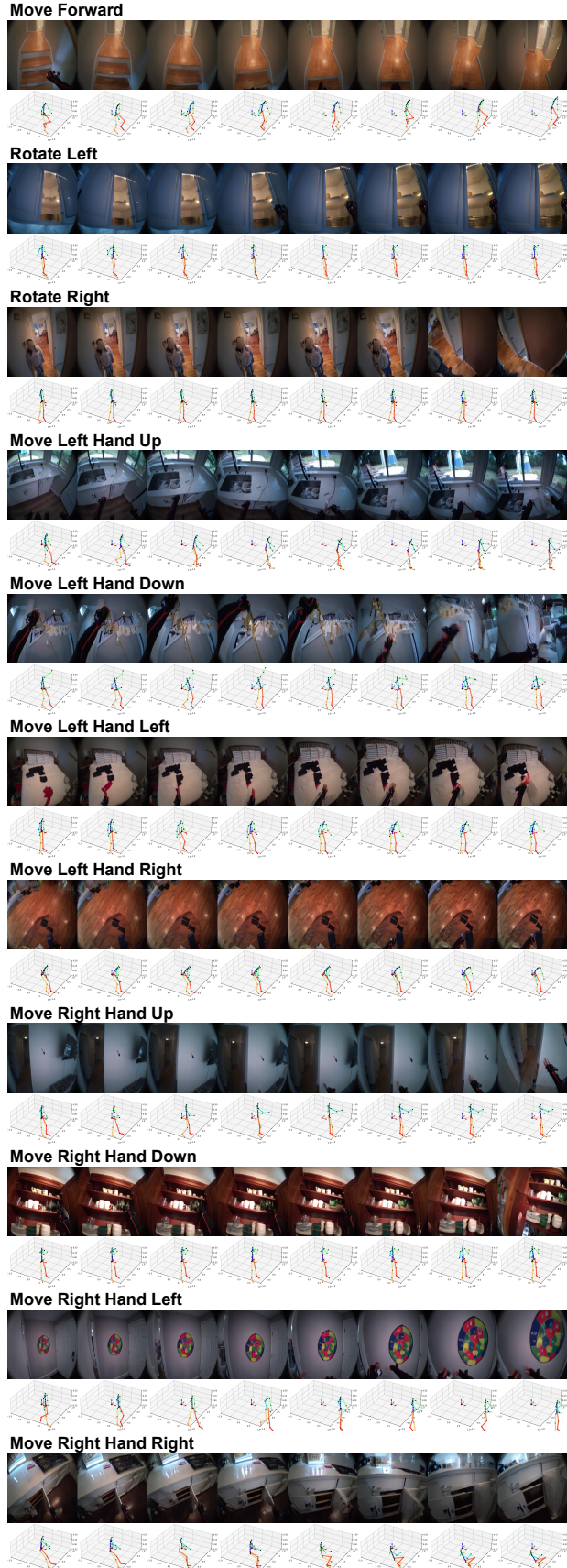


Figure 8: **Atom Actions Generation.** We include video generation examples of different atomic actions specified by 3D-body poses.

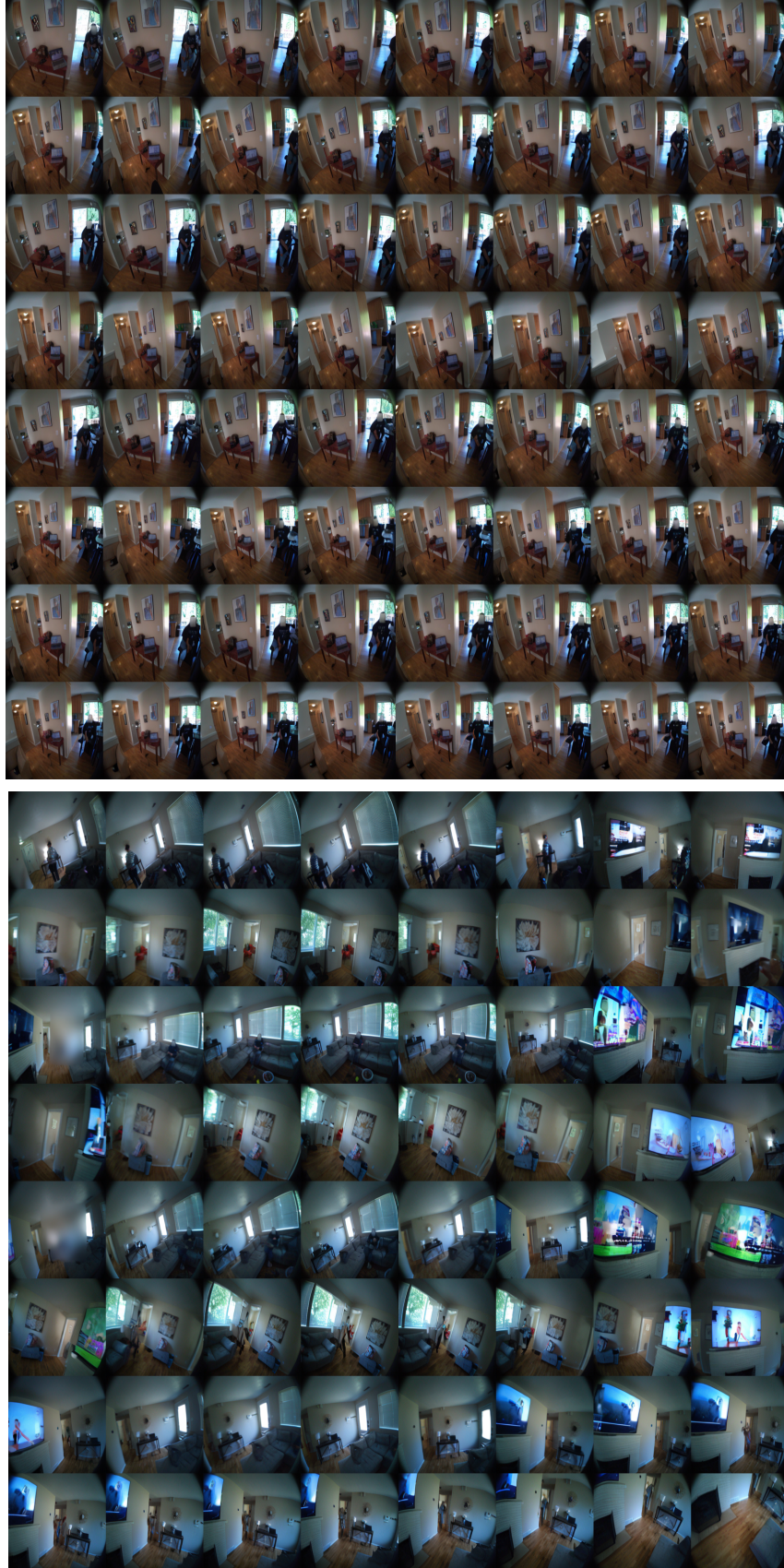


Figure 9: **Generation Over Long-Horizons.** We include 16-second video generation examples.

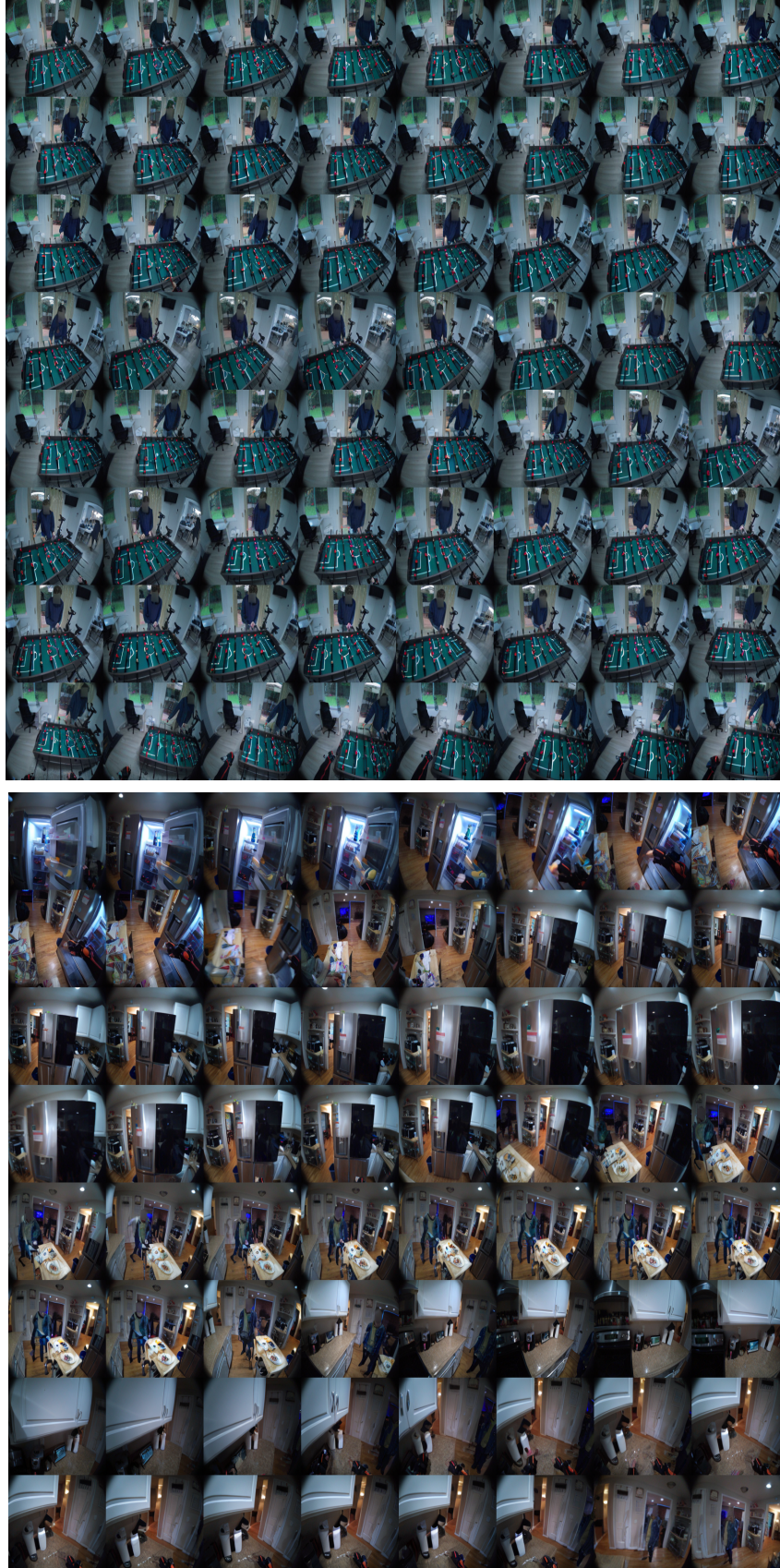


Figure 10: **Generation Over Long-Horizons.** We include 16-second video generation examples.



Figure 11: **Generation Over Long-Horizons.** We include 16-second video generation examples.

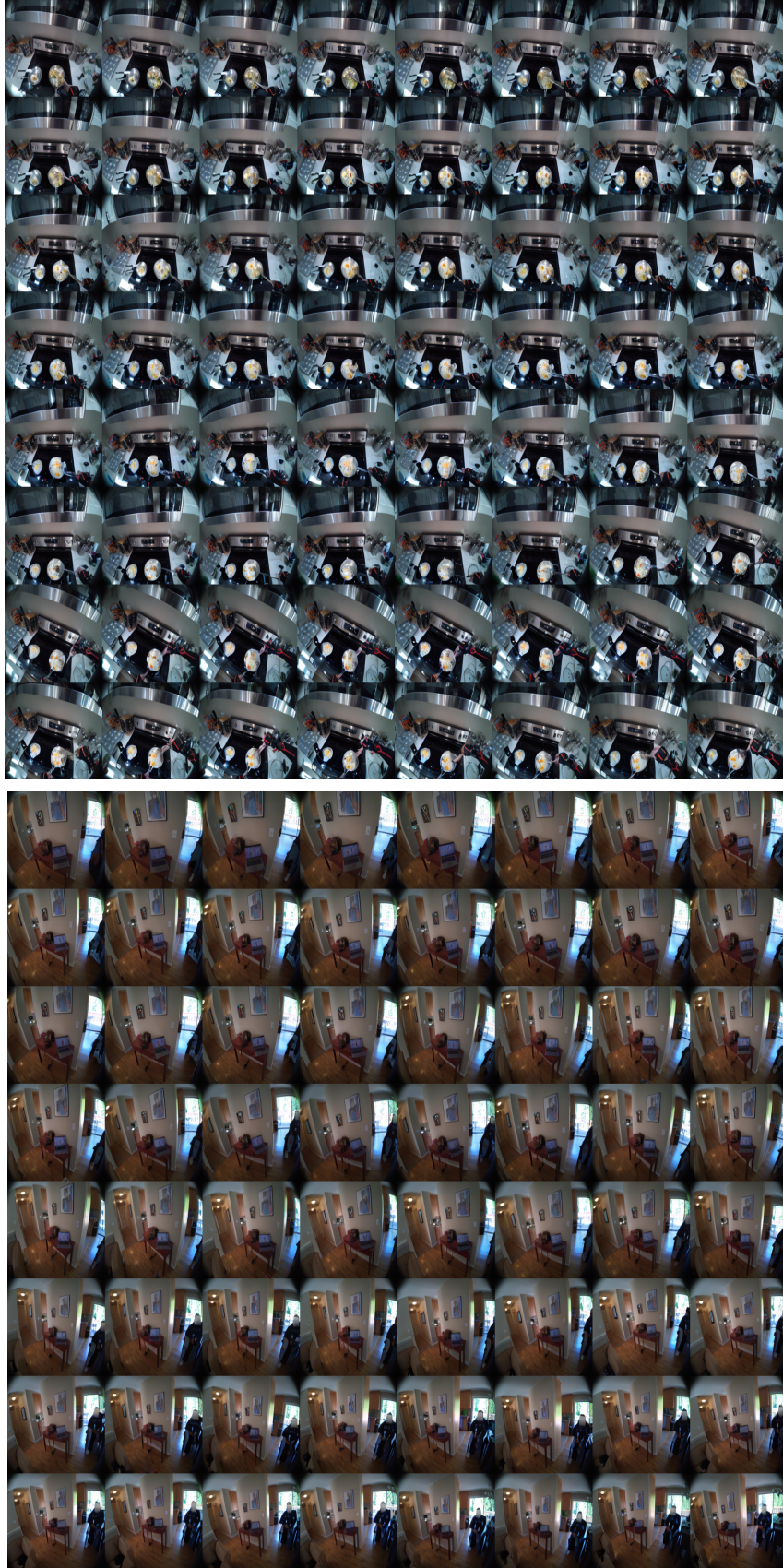


Figure 12: **Generation Over Long-Horizons.** We include 16-second video generation examples.

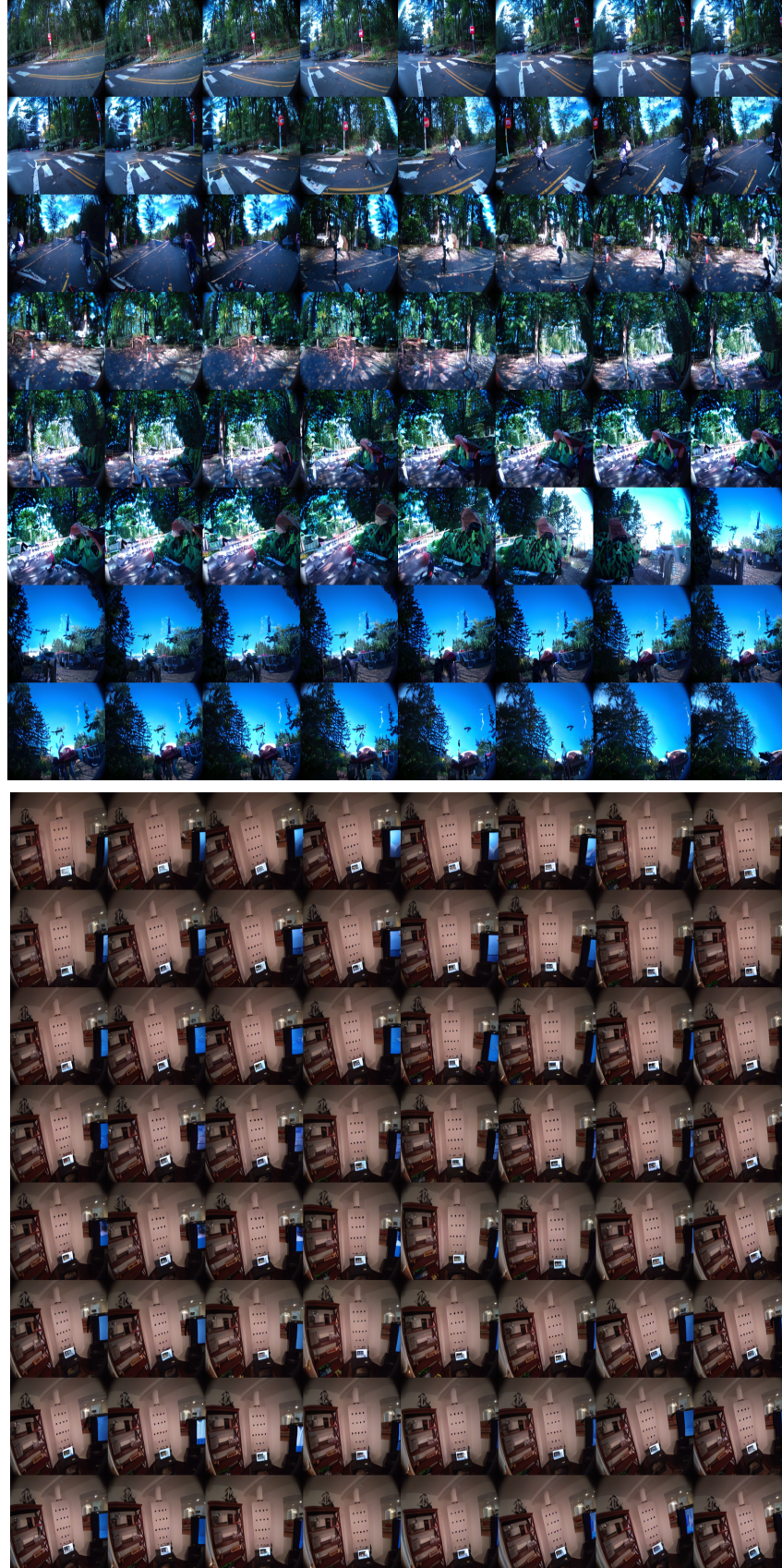


Figure 13: **Generation Over Long-Horizons.** We include 16-second video generation examples.



Figure 14: **Generation Over Long-Horizons.** We include 16-second video generation examples.

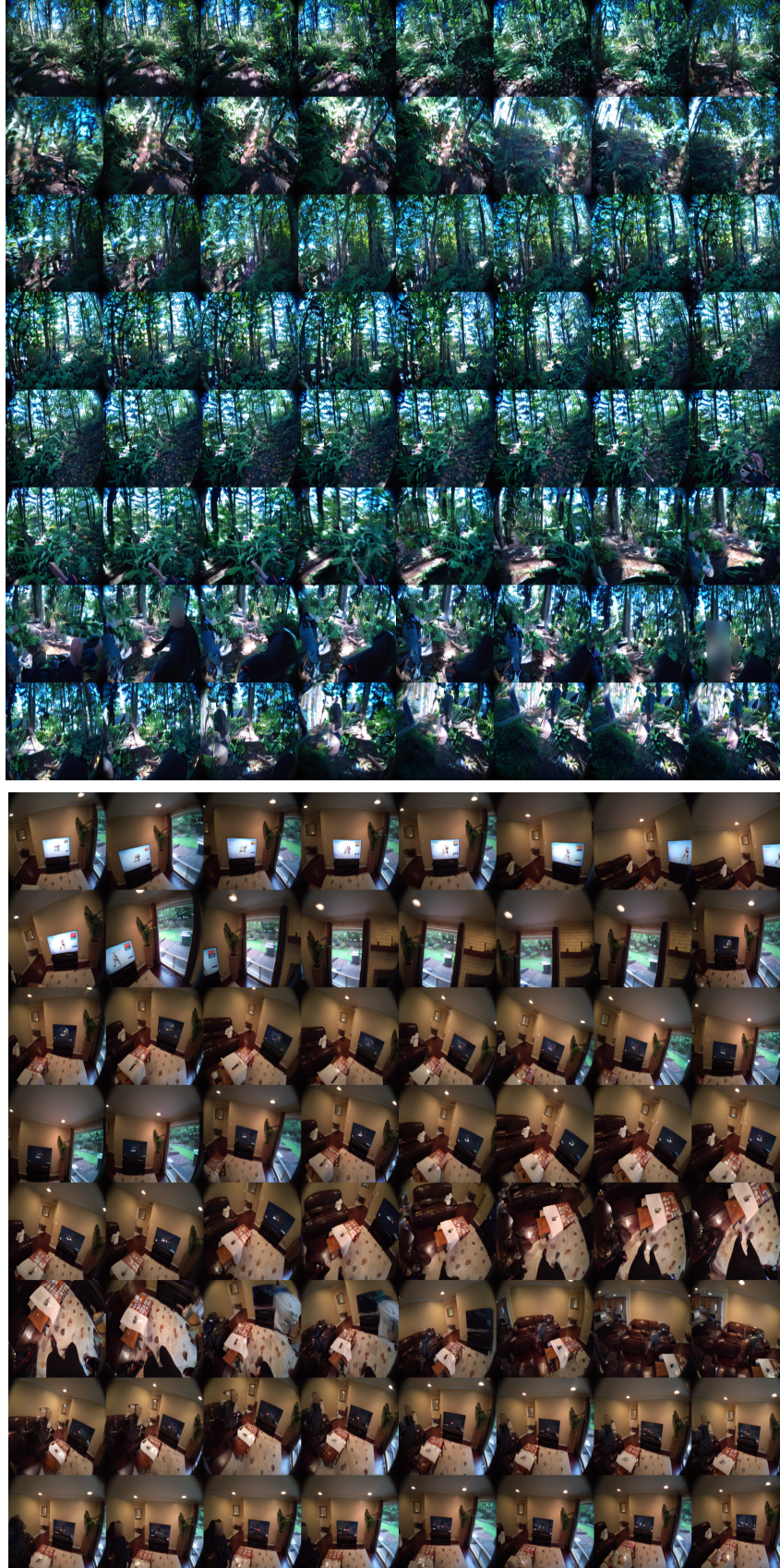


Figure 15: **Generation Over Long-Horizons.** We include 16-second video generation examples.

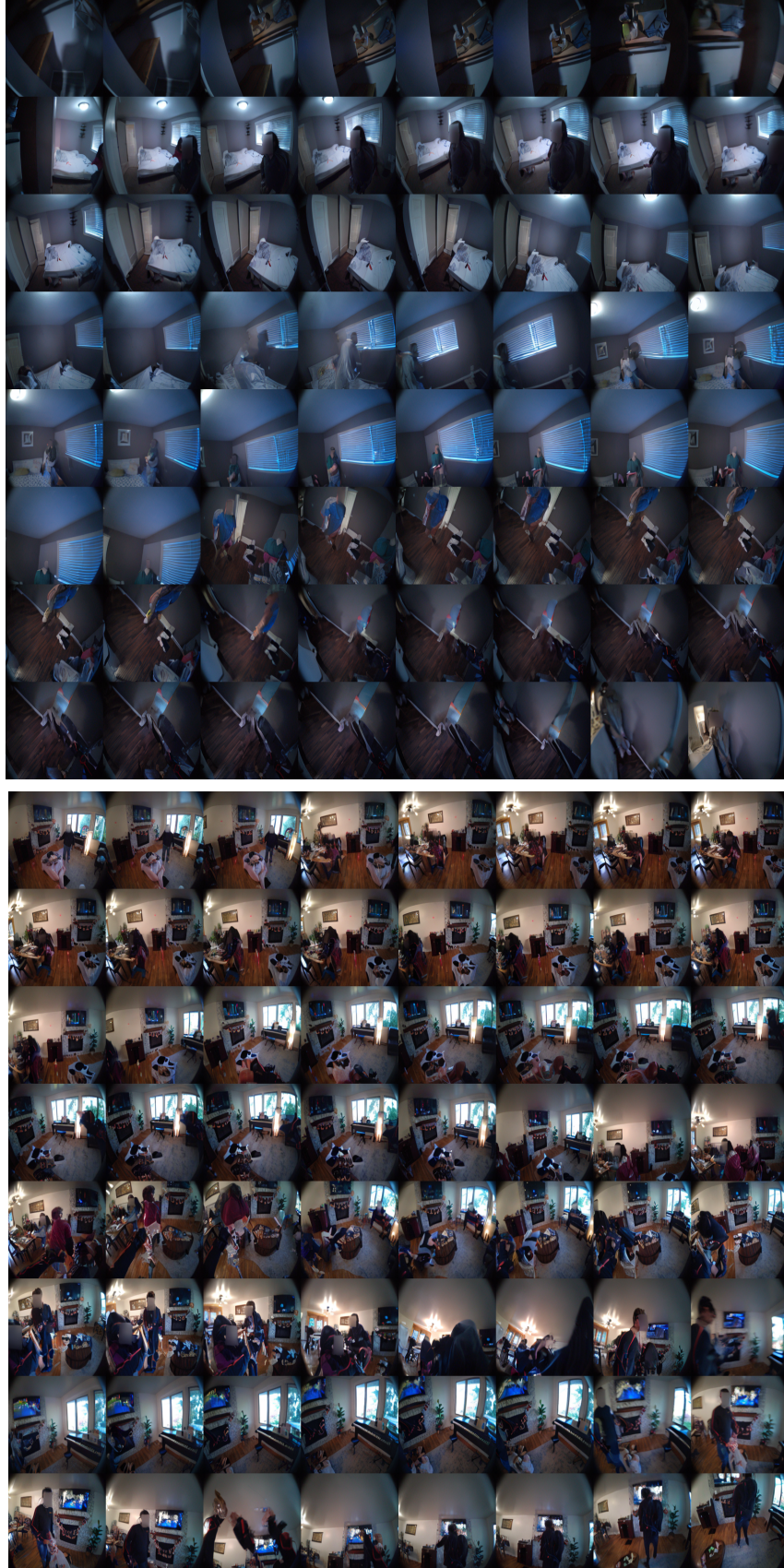


Figure 16: **Generation Over Long-Horizons.** We include 16-second video generation examples.



Figure 17: **Generation Over Long-Horizons.** We include 16-second video generation examples.

36 **References**

- 37 Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. *arXiv*
38 *preprint arXiv:2412.03572*, 2024.
- 39 Reuven Y Rubinstein. Optimization of computer simulation models with rare events. *European Journal of*
40 *Operational Research*, 99(1):89–112, 1997.
- 41 Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness
42 of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and*
43 *pattern recognition*, pages 586–595, 2018.